

Criteria for Consciousness in Artificial Intelligent Agents

Raúl Arrabales
Universidad Carlos III de Madrid
Avda. Universidad, 30
28911 Leganés. Spain
+34 91 6249111
rarrabal@inf.uc3m.es

Agapito Ledezma
Universidad Carlos III de Madrid
Avda. Universidad, 30
28911 Leganés. Spain
+34 91 6249110
Ledezma@inf.uc3m.es

Araceli Sanchis
Universidad Carlos III de Madrid
Avda. Universidad, 30
28911 Leganés. Spain
+34 91 6249423
masm@inf.uc3m.es

ABSTRACT

Accurately testing for consciousness is still an unsolved problem when applied to humans and other mammals. The inherent subjective nature of conscious experience makes it virtually unreachable to classic empirical approaches. Therefore, alternative strategies based on behavior analysis and neurobiological studies are being developed in order to determine the level of consciousness of biological organisms. However, these methods cannot be directly applied to artificial systems. In this paper we propose both a taxonomy and some functional criteria that can be used to assess the level of consciousness of an artificial intelligent agent. Furthermore, a list of measurable levels of artificial consciousness, *ConsScale*, is defined as a tool to determine the potential level of consciousness of an agent. Both the mapping of consciousness to AI and the role of consciousness in cognition are controversial and unsolved questions, in this paper we aim to approach these issues with the notions of *I-Consciousness* and embodied intelligence.

Categories and Subject Descriptors

I.2 [Artificial Intelligence]: General – *cognitive simulation, philosophical foundations*.

General Terms

Design, Theory.

Keywords

Machine Consciousness, Artificial Consciousness, cognitive agents, Cognitive Modeling.

1. INTRODUCTION

Determining the level of consciousness of a living organism is a hard problem. One could think that some sort of Turing test might be a plausible solution [31]. It is indeed what we do everyday when we perceive other subjects as conscious beings. These kinds of test that we are used to perform unconsciously are based on verbal report and observed behavior. We perceive other humans acting as if they were conscious and thus we infer they actually are. However, we do not have any scientific proof that others experiment any subjective life because we cannot perceive it directly [15]. Therefore, from a pure scientific point of view, zombies (organisms behaving as conscious beings but without any

inner feeling) are conceivable, although probably not possible [7].

Some approaches have been proposed in order to overcome the issue of scientific proof of consciousness. From a philosophical standpoint, Dennett has proposed the *heterophenomenology* method, which consists on the application of the scientific method to both third-person behavior analysis and first-person self report [10]. From the neuroscience perspective, Seth, Baars and Edelman propose a set of criteria for consciousness in humans and other mammals [29]. A number of these criteria are based on neurobiological aspects. If the neuronal structures and the associated activity pattern that gives place to consciousness are identified, then we can look for them in animals endowed with a central nervous system [5]. Analogously, if some behavior patterns are identified as uniquely produced by a conscious subject, we can design experiments where these behaviors are tested. However, when it comes to artificial agents, most of the assumptions mentioned above cannot be directly applied. The following are the main reasons why we think the former criteria should not be used for evaluating artificial agents:

- **Artificial agents have different underlying machinery.** At the biological level, behavior of mammals is controlled by the endocrine and nervous systems. Even though some artificial agents are inspired or try to simulate the biological nervous system, their design is quite far from a realistic emulation. Therefore, it does not make sense, for instance, to look for a strong connection between thalamus and cortex as a possible sign of an underlying mechanism for consciousness in an artificial implementation (given the case that the implementation under study is endowed with a simulated thalamocortical structure).
- **Artificial agent's behavior produces different patterns.** Moving the observer's point of view from the biological level to the behavioral level, human behavior can be seen as regulated by cultural rules. Human ontogeny gives place to different behavioral patterns as a subject develops situated in a cultural environment. Given that the development of artificial agents differs from that, their behavior should not be analyzed following the same criteria that are applied to humans.
- **Lack of verbal report.** This is one of the key differences between human's behavior and artificial agents' behavior. Accurate verbal report (AVR) is probably the main way we can find out about the inner life experienced by a human subject. Given the lack of this kind of communication skills in artificial systems, AVR as we know it cannot be used to evaluate artificial agents.

Taking into account the reasons mentioned above and the fact that human culture strongly determine the production of consciousness in humans [6], we argue that the kind of consciousness that could be potentially produced in artificial agents would be of a different nature (although we think it still could be called consciousness, i.e. *machine consciousness* or *artificial consciousness*). Consequently, we believe that criteria for machine consciousness should be studied from the perspective of a specifically defined taxonomy of artificial agents. Even though some of the classes of artificial agents defined in this taxonomy cannot be directly compared with a corresponding example of biological organisms, both biological phylogenetic and human ontogenetic analogies can often be used to better understand the level of consciousness that can be associated to a particular class of agents, e.g. [16].

In the next section we aim to provide a comprehensive description of the main aspects of consciousness and their basic roles in cognition. Additionally, we redefine the dimensions of consciousness in terms of artificial intelligent agents, and therefore we characterize machine consciousness by analyzing the fundamental building blocks required in an agent's architecture in order to produce the functionality associated with consciousness. Subsequently, in section 3, we discuss key particular functions of consciousness and their interaction in agent's cognitive processes. In section 4, we have taken into account both the key functions of consciousness and agent's basic architectural features to propose a taxonomy for artificial agents, where a concrete level of machine consciousness is assigned to each agent category. Section 5 provides a framework for classifying agents under the light of the proposed taxonomy. Finally, we conclude in section 6 with a brief discussion of current state of the art in terms of our proposed taxonomy.

2. CHARACTERIZING MACHINE CONSCIOUSNESS

Setting aside the discussion about whether or not a categorical implementation of an artificial form of consciousness is possible, we have adopted an incremental approach in which we consider that certain aspects of consciousness can be successfully modeled in artificial agents; while other aspects might be still out of the reach given the current state of the art in the field of machine consciousness. In this scenario, we need to define which are the conceptual building blocks integrated in a possible machine consciousness implementation. Then we could test the presence of these functional components and their interrelation within a given system in order to assess its potential level of machine consciousness. However, the definition of these components would require a complete understanding of 'natural' consciousness, and given that the quest for consciousness has not yet come to a successful end, a more modest framework has to be established in the realm of artificial systems. But, what are the components of consciousness that we are not able to explain or concretely define so far? We need to decompose, or at least conceptually decouple consciousness dimensions in order to be able to answer this question.

2.1 The Dimensions of Consciousness

An extremely complex phenomenon like consciousness can be seen as a whole, or more conveniently, it can be analyzed as if it was composed of two interrelated dimensions. A conceptual division can be outlined when a distinction is made between

phenomenology and access [6]. While the access dimension (*A-Consciousness*) refers to the accessibility of mind contents for conscious reasoning and volition, phenomenology (*P-Consciousness*) is related to the subjective experience or *qualia*, i.e. how does it feel to be thinking about something, or what is it like to be someone else, as Nagel would formulate it [20]. Understanding how P-Consciousness is produced by biological organisms is a controversial problem usually regarded as the *explanatory gap* [17], which still remains to be closed (if ever possible). While the access dimension of consciousness has an obvious function, namely guiding conscious reason and action; the phenomenal dimension lacks a generally accepted function (see [6] and [8] for a detailed discussion on the matter). Qualia could be just a side effect produced by access mechanisms, or it could play a key role in the integration of multimodal perception [27]. What is generally accepted is that rather than binary properties, both access and phenomenal aspects of consciousness come in various degrees. Therefore, we think it is possible to represent a range of degrees of consciousness in a bi-dimensional space defined by phenomenal and access dimensions (see Figure 1). The access dimension represents the informational aspect of consciousness, while the phenomenal dimension represents the associated '*what-is-it-like-ness*'.

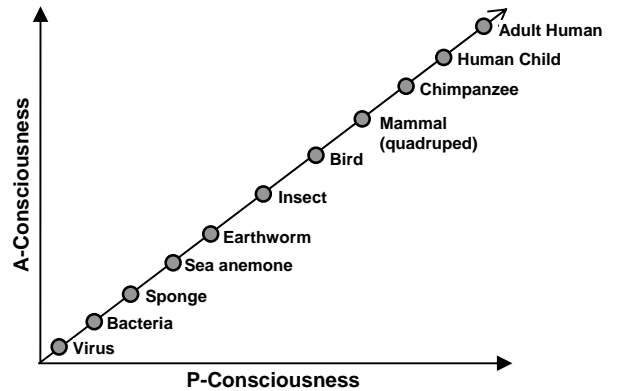


Figure 1. Consciousness bi-dimensional space in biological phylogenics.

The questions of having A-Consciousness without P-Consciousness and *vice versa* are typically controversial issues in the study of consciousness. In the present work, we have adopted the assumption that machine consciousness and 'biological' consciousness are actually different phenomena. Therefore, different kinds of consciousness could be present in artificial agents, and these new versions of machine consciousness could follow different rules in terms of the conceptual link between A-Consciousness and P-Consciousness. While we assume that both A-Consciousness and P-Consciousness increase uniformly at the same rate in biological phylogeny (as depicted in Figure 1), we consider that all combinations are *a priori* possible in artificial agents. We believe that the evolutionary forces involved in the design of biological organisms have always produced functionally coherent machinery; hence zombies or *P-Unconscious* (A-Consciousness without P-Consciousness) and *A-Unconscious* (P-Consciousness without A-Consciousness) do not naturally exist. Nevertheless, there exist cases of individuals that after suffering cerebral vascular accidents or traumatic brain injury become P-Unconscious or A-Unconscious in some respects and degrees. For

instance, brain-injured patients who have developed prosopagnosia are unable to consciously recognize faces despite being able to recognize any other visual stimuli. Even though prosopagnosic patients are unable to experience any feeling of familiarity at the view of faces of their closest relatives (loss of P-Consciousness), other cognitive operations are still performed with the perceived faces – a covert face recognition takes place – but their output fails to reach consciousness (a disorder of A-Consciousness). However, some A-Consciousness capability remains in many patients as they are usually able to implicitly access to knowledge derived from ‘P-Unconsciously’ unrecognized faces [28].

It is also important to distinguish between consciousness as it is applied to creatures and consciousness as it is applied to mental states [19]. Essentially, a conscious subject can have conscious and unconscious mental states. In the prosopagnosia example discussed above, conscious individuals fail to have P-Consciousness of faces at view and their A-Consciousness is also impaired to that respect. However, these subjects can perfectly be A-Conscious and P-Conscious of the voice and speech of their relatives or any other person. In this paper, we generally refer to creature consciousness, hence evaluating the potential level of consciousness of individuals as per their ability to have P-Conscious and A-Conscious states. The particular contents of the mental states will be analyzed later as part of the method to establish a taxonomy for machine consciousness.

2.2 A Computational Approach to Consciousness in Intelligent Agents

The possible functionality of P-Consciousness and the possibility of effectively having one dimension of consciousness without another remain unanswered questions. Therefore, the interrelation between access and phenomenology remains highly unclear and controversial. Some authors even consider P-Consciousness as an epiphenomenal process, hence independent of behavior (see for instance [32], while others tend to identify a key functional role in qualia [21].

Following a pure computational approach we could consider both A-Consciousness and P-Consciousness as being the same functional process, thus neglecting the possibility of subjective experience in artificial agents. However, we think that a different dimensional decomposition is to be made in the realm of machine consciousness (see Figure 2). Although the nature and required underlying machinery for qualia are not known, we believe that some functional characterization of P-Consciousness can be made. Therefore, we have adopted a functional point of view, in which we introduce a redefined dimension of consciousness called *Integrative Consciousness* (I-Consciousness). In our conception of machine consciousness, we have taken the assumption that I-Consciousness represents the functional aspect of P-Consciousness that exists in conscious biological organisms.

In order to characterize consciousness as a property of agents we need to formally define the basic components of an artificial situated agent. Such an agent interacts with the environment by retrieving information both from its own body and from its surroundings, processing it, and acting accordingly. Following Wooldridge’s definition of abstract architectures for intelligent agents [33], and taking into account the embodiment aspect of situated agents, we have identified a set of essential architectural

modules: sensors, sensorimotor coordination, internal state, and effectors. These modules implement the following processes: perception, reason, and action. Consequently, the following abstract architectural components can be identified:

- **Body (*B*)**. Embodiment is a key feature of a situated agent [11]. Agent’s body can be physical or software simulated (as well as its environment). A boundary is established between agent’s body and its environment (*E*). The rest of components are usually located within this boundary. We believe that it is important to make a distinction between agent’s body (or *plant* if we take a control theory standpoint) and the environment, as the first is directly controlled while the latter is indirectly controlled. The definition of the body of an agent is important as it determines what sensors it can use, how its effectors work, and ultimately how its perception and behavior is affected by its physical embodiment. Owning an active body is essential for the acquisition of consciousness.
- **Sensory Machinery (*S*)**. Agent’s sensors are in charge of retrieving information from the environment (exteroceptive sensors) or from the agent’s own body (proprioceptive sensors).
- **Action Machinery (*A*)**. In order to interact with the environment the agent uses its effectors. Agent’s behavior is composed of the actions ultimately performed by this machinery.
- **Sensorimotor Coordination Machinery (*R*)**. From purely reactive agents to deliberative ones, the sensorimotor coordination module is in charge of producing a concrete behavior as a function of both external stimuli and internal agent’s state.
- **Memory (*M*)**. Internal agent’s state is represented both by its own structure and stored information. Memory is the mean to store both perceived information and new generated knowledge. We consider that even agents that do not maintain state can be said to have a minimal state represented by its own structure, i.e. preprogrammed sensorimotor coordination rules.

As Wooldridge has pointed out [33], different classes of agents could be obtained depending on the concrete implementation of the abstract architecture. Following the notation that we have adopted, we could say that different sensorimotor coordination functions give place to different classes of agents. For instance, reactive agents or BDI agents [23]. While sensorimotor coordination of reactive agents is characterized by a direct mapping from situation to action, BDI agents decision making is based on internal state representing beliefs, desires, and intentions.

In computational terms, consciousness can be regarded as a unique sequential thread that integrates concurrent multimodal sensory information and coordinates voluntary action. Hence, consciousness is closely related with sensorimotor coordination. Our aim is to establish a classification of agents according to the realization of the functions of consciousness in the framework of agent’s sensorimotor coordination.

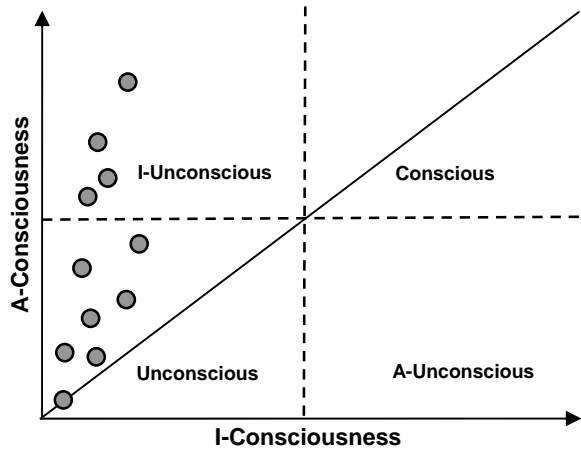


Figure 2. Machine Consciousness bi-dimensional space.

According to the Global Workspace Theory [4], loads of information are acquired by the senses continuously, and many interim coalitions of specialized processors run concurrently collaborating and competing for space in the working memory, which is the arena where the serial mechanism of attention selects the contents that will be conscious at any given time. In this scenario, A-Consciousness refers to the accessibility of contents for their usage in conscious processing. In accordance with the Global Access Hypothesis [3], the output of unconscious processors, like for instance a face recognition module, can be accessed by other processors, and be finally used to form the conscious contents of the mind. Baars argues that the aggregation of processors is produced by the application of contexts. However, access is not the only feature that is required to form a conscious experience. Coherent context criteria need to be selected and applied adaptively. We argue that I-Consciousness is the mechanism that allows the formation of *coherent contents* of consciousness.

A coherent content of consciousness is one that provides a desired functionality which successfully adapts to current environment situation. For example, given the access to the recognition of a face, a conscious content should be formed including a feeling of familiarity (or a familiarity flag setting aside the phenomenal dimension) if the face belongs to a known person. This is a desired functionality for a social agent, and the access property alone cannot provide it. Basically, we argue that I-Conscious dimension of machine consciousness represents the functionality that caused qualia to be selected by evolution in biological organisms.

3. FUNCTIONS OF CONSCIOUSNESS

As mentioned above, the question of what do qualia do in biological organisms is a controversial one. In this paper we propose that a naturalistic approach on the origin of consciousness can be applied to machine consciousness, and therefore identify the functions that can render an agent conscious (in the sense of Artificial Consciousness). In a vast ocean of information where A-Consciousness provides access to virtually any content of the agent's mind, I-Consciousness provides the mechanism for the emergence of a unique and coherent story out of the chaos. This story is the stream of consciousness, the metaphorical movie that is playing within our heads. As Dennett has pointed out [10], the

process of making this narrative could be based on a kind of pandemonium, where different narrative versions suffer reiterative edition and review until they are presented as the official published content of the mind, i.e. they become conscious contents of the mind.

In order to determine the functionality that has to be included as part of I-Consciousness we have analyzed the very basic functions that need to be considered in the making of a story out of sensory information. Note that different functions can be considered depending on the problem domain, agent physical capabilities, and internal state representation richness. In fact, each specific class of organism is designed to perceive different realities from the world, thus limiting what can be available to consciousness. For instance, while some animals (including humans) have the ability to perceive social relations, other animals endowed with similar senses are unable to internally represent such complex percepts.

In this work, we have adopted the assumption that single modality percepts acquired by the agent are combined using contextualization in order to form complex multimodal percepts [2]. Understanding how this process is performed in the brain, subsequently giving place to a unique version (or story) of conscious perception is known as the *binding problem* [25]. From a machine consciousness perspective, the binding problem is solved functionally by applying a contextualization mechanism. This contextualization process alone can generate multiple complex percepts. However, it is the combination of A-Consciousness and I-Consciousness which permits the construction of coherent and adaptive complex percepts. The set of finally accepted percepts form a unique and coherent stream of consciousness, which the agent exploits to develop other higher level cognitive functions.

Out of the set of cognitive functions that an intelligent agent could potentially exhibit, the following list of functions specifically characterizes the behavior of a conscious agent: Theory of Mind (ToM) and Executive Function (EF). ToM is the ability to attribute mental states to oneself and others. From a human developmental standpoint, Lewis suggests four stages in the acquisition of ToM: (1) "I know", (2) "I know I know", (3) "I know you know", and finally (4) "I know you know I know" [18]. The term EF includes all the processes responsible for higher level action control, in particular those that are necessary for maintaining a mentally specified goal and for implementing that goal in the face of distracting alternatives [22]. Attention is an essential feature of EF. It represents the ability of the agent to direct its perception and action, i.e. selecting the contents of the working memory out of the entire mind's accessible content. Planning, coordination, and set shifting (the ability to move back and forth between tasks) are also key processes included in EF. We argue that the integration of all of these cognitive functions could build an artificial conscious mind. However, each of the mentioned functions could also be implemented independently or partly integrated with other cognitive functions, thus giving place to different levels of implementation of artificial consciousness as discussed in the next section.

4. LEVELS OF MACHINE CONSCIOUSNESS

Table 1 describes *ConsScale*, which is a list of potential levels of consciousness for artificial agents. This scale has been defined in terms of reference agent abstract architectures and characteristic behaviors. The characteristic behavior assigned to each level has been derived from the functionality of consciousness discussed above. As illustrative analogy, machine consciousness levels are assigned a comparable level of consciousness in biological phylogenics and human ontogeny.



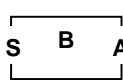
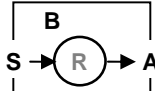
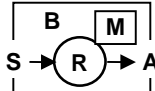
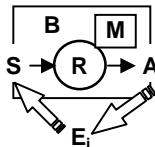
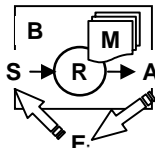
The first level in the scale, *Disembodied*, refers to a ‘proto-agent’ and serves as an initial reference that remarks the importance of a defined body as a requirement for defining a situated agent. The rest of the scale comprises a set of twelve ranks, where lower levels are subsumed by higher ones. Therefore, each stage of the incremental development of an artificial agent could be identified by a concrete level. Levels 0 and 1, *Isolated* and *Decontrolled* respectively, are also conceptual references which help characterize situatedness in terms of the relation with the environment. Both classes represent inert bodies lacking any functionality or interaction with the medium except the inevitable derived from the physical properties of their inactive bodies.

Therefore, these classes cannot be defined as situated agents. Level 2, *Reactive*, defines a classical reactive agent which lacks any explicit memory or learning capabilities. From level 2 onwards the agents make use of the environment as the mean to close the feedback loop between action and perception. Hence, all agent types above level 1 can be regarded as situated agents.

Although we are explicitly focusing in individual agent evaluation, it is important to note that additional learning or adaptation processes could exist at an evolutionary plane (assuming that agents are able to replicate, mutate, and evolve). For instance, although reactive rules are fixed for a level 2 individual, adaptation of reactive responses in a population of agents could take place over the generations.

Level 3, *Rational*, can be identified as the simplest form of a classical deliberative agent. At this level, the agent’s internal state is maintained by a memory system and sensorimotor coordination is a function of both perceived and remembered information. Proprioceptive sensing can be present at this level; however, it is not producing any self-awareness. The next level, *Attentional*, is characterized by an attention mechanism, which allow the agent to select specific contents both from the sensed and stored state information.

Table 1. Artificial Agents Consciousness Scale (*ConsScale*)

Level of Machine Consciousness	Agent Architecture	Short Description	Characteristic Behavior	Biological Phylogeny	Human Ontogeny
Level -1 <i>Disembodied</i>	E 	Boundaries of the agent are not well defined. It can be confounded with the environment.	None. It is not a situated agent.	Amino acid as part of a protein.	n/a
Level 0 <i>Isolated</i>	E 	Obvious distinction between body and environment, but no autonomous processing.	None. It is not a situated agent.	Isolated chromosome.	n/a
Level 1 <i>Decontrolled</i>	E 	Presence of sensors and/or actuators, but no relation between them.	None. It is not a situated agent.	Dead bacteria	n/a
Level 2 <i>Reactive</i>	E 	Fixed reactive responses. R establishes an output of A as a predetermined function of S.	No higher function. Primitive situatedness based on reflexes.	Virus	n/a
Level 3 <i>Rational</i>	E 	Actions are a dynamic function of both memory and current information acquired by S.	Basic ability to learn and proprioceptive sensing allow orientation and positioning behavior.	Earthworm	1 Month.
Level 4 <i>Attentional</i>	E 	Attention mechanism selects E _i contents from S and M. Primitive emotions.	Ability to direct attention toward selected E _i allows attack and escape behaviors.	Fish	5 Months.
Level 5 <i>Executive</i>	E 	Multiple goals can be interleaved as they are explicitly represented in memory.	Set shifting capability allows multiple goal achievement. Basic emotional learning.	Quadruped mammal	9 Months.

Level 6 <i>Emotional</i>		Stable and complex emotions. Support for ToM stage 1: “I know”.	Complex emotions provide a self-status assessment and influence behavior.	Monkey	1 Year.
Level 7 <i>Self-Conscious</i>		Support for ToM stage 2: “I know I know”.	Self-reference makes possible advanced planning. Use of tools.	Monkey	1.5 Years.
Level 8 <i>Empathic</i>		Support for ToM stage 3: “I know you know”.	Making of tools.	Chimpanzee	2 Years.
Level 9 <i>Social</i>		Support for ToM stage 4: “I know you know I know”.	Linguistic capabilities. Able to develop a culture.	Human	4 Years.
Level 10 <i>Human-Like</i>		Human like consciousness. Adapted Environment (E_c).	Accurate verbal report. Behavior modulated by culture (E_c).	Human	Adult
Level 11 <i>Super-Conscious</i>		Several streams of consciousness in one self.	Ability to synchronize and coordinate several streams of consciousness.	n/a	n/a

A level 5 agent, *Executive*, includes a more complex internal state representation, which provides set shifting capabilities. The achievement of multiple goals is sought thanks to a higher coordination mechanism that shifts attention from one task to another. Level 6, *Emotional*, is the first level in which an agent can be to certain extent regarded as conscious in the sense of self-awareness. The main characteristic of this level is the support for ToM stage 1, “I know”. Complex emotions are built as a combination of basic emotions and they are not only used to evaluate external objects but to assess the internal agent status. Level 7, *Self-Conscious*, corresponds to the emergence of self-consciousness. At this level the agent is able to develop higher order thoughts [26], i.e. thoughts about thoughts, and more specifically thoughts about itself. Consequently it presents support for ToM stage 2, “I know I know”. Progressing to the next level, *Empathic*, the internal representation of the agent is enriched by *inter-subjectivity*. In addition to the model of the self, others are also seen as selves; hence, they are consequently assigned a model of subjectivity. This is the seed for a complex

social interaction. The next step is represented by level 9, *Social*, where ToM is fully supported. Level 10, *Human-Like*, represents the sort of agent that is endowed with the same level of consciousness as a healthy adult human has. Therefore, the formation of a complex culture is a feature of this level. Finally, level 11 or *Super-Conscious*, refers to a kind of agent able to internally manage several streams of consciousness, while coordinating a single body and physical attention. A mechanism for coordination between the streams and synchronized access to physical resources would be required at this level.

5. CLASSIFYING AGENTS USING *ConsScale*

The levels of artificial consciousness defined in *ConsScale* are characterized by abstract architectural components and agent’s behavior. The architecture components represent functional modules whose integration makes possible the emergence of a characteristic behavior. Therefore, at least one behavior-based test can be associated to each level in order to assess if a particular

agent fulfills the minimum required behavioral pattern for that level. In fact, an agent can only be assigned a concrete level if and only if it is able to show the behavioral pattern of that level as well as the behavioral patterns of all lower levels, e.g. even though an agent is able to pass *ConsScale* level 7 behavior test, it does not necessarily imply that it can be regarded as Self-Conscious in terms of *ConsScale*. It would also need to comply with all lower levels.

As discussed above, the three first reference levels (*Disembodied*, *Isolated*, and *Decontrolled*) are a special case as they do not actually describe situated agents. Therefore, there are no behavioral tests associated to any of these first three levels. A given agent could be assigned either of these initial reference levels just by analyzing its architectural components. In contrast, from level 2 onwards a characteristic behavior pattern is defined per *ConsScale* level. This characteristic pattern should be taken as the base of any behavior test that can be assigned to a particular level. Reference behavior patterns for levels 2 to 11 are discussed below.

The characteristic behavior of level 2, *Reactive*, is the reflex, hence an agent able to autonomously react to any given environment situation is said to comply with level 2. When the response to a given environment state is not fixed, but it is a function of both the information acquired by S and agent's internal state, then the agent is said to comply with level 3, *Rational* (note that some proprioceptive sensing mechanism is required to make agent's internal state available in R, so it can be an input of the sensorimotor coordination function). Most BDI-type agents ([23]) could be classified as level 3 in terms of *ConsScale*.

If the agent is able to direct attention to a selected subset of the environment state (E_i) while other environmental variables are also sensed but ignored in R, and the selected perception is evaluated in terms of agent's goals so subsequent responses are adapted (primitive emotions), then the agent is said to comply with level 4, *Attentional*. Level 4 agents are able to show specific attack or escape behaviors and trial and error learning. The ability to pay attention toward specific objects or events gives place to the formation of directed behavior, i.e. agent can develop behaviors clearly related to specific targets, like following or running away. Additionally, level 4 agents can have primitive emotion mechanisms in the sense that the objects to which attention is paid are elementally evaluated as positive or negative. A positive emotion triggers decrease of distance behavior or bonding to selected object, while negative emotion triggers increase of distance and reinforcement of boundaries toward selected object [9].

If an agent that can be successfully classified as *Attentional* in terms of *ConsScale* also exhibits set shifting and basic emotional learning capabilities, then it can be regarded as *Executive* (*ConsScale* level 5). In addition to advanced planning, emotional learning is another characteristic that can be observed in some degree at this level, as the most emotionally rewarding tasks are assigned more time and effort.

By basic emotional learning we mean that the agent is able to learn basic rules from one task and adapt its behavior consequently in the performance of that particular task. In contrast, *Emotional* (*ConsScale* level 6) agents are characterized by complex emotions and complex emotional learning. This

means that the agent generalizes the learned lessons to its general behavior, furthermore, emotions are also assigned to the self and self-status monitoring and evaluation gives place to a sense of "I know" (support for ToM stage 1). Even though a representation of the self is considered as an input of the sensorimotor coordination function, this is an implicit symbol. However, level 7 (*Self-Conscious*) is described by an explicit symbol for the self, which enables self-recognition. The reference behavior test for this level would be the mirror test, which although originally applied to primates [13], has also been adapted to other mammals and even artificial agents. Takeno et al. have proposed a specific experiment design to test whether a robot is able to recognize its own image reflected in a mirror [30]. Planning capabilities are extended as the self is integrated both in the current state representation and future state estimation. Behavior at this level is also illustrated by the ability to use tools (see for instance [1]).

ConsScale Level 8 (*Empathic*) is achieved by an agent when it shows that it maintains a model of others, and therefore it collaborates accordingly with other agents in the pursuit of a common goal. In fact, joint goals require this, and the need for socially aware plans in BDI agents has been considered some time ago [24].

In level 9, *Social*, the internal model of other selves is enhanced with a full support of ToM. This means that characteristic behavior of this level is defined by sophisticated Machiavellian strategies (or social intelligence) involving social behaviors like lying, cunning, and leadership. In other words, an agent A could be aware that another agent B could be aware of A's beliefs, intentions, and desires. Advanced communication skills are the characterization of this level behavior, where, for the first time, an agent would be able to purposely tell lies. There exist mathematical models of the dynamics of Machiavellian intelligence that could be used to test these sort of behaviors with artificial agents [14].

While, the obvious test for level 10, *Human-Like*, is the Turing test [31], also accurate communications skills (language) and the creation of a culture would be a clear feature of level 10. Other key characteristics are that the agent is able to profoundly modify its environment and society. The fluidity between social and technical intelligence permits the extension of its own knowledge using external media (like written communication) and technological advances are also possible.

Finally, we cannot envisage any conclusive behavior test for level 11 due to the lack of known exemplifying references.

6. CONCLUSIONS

We have proposed *ConsScale* as a machine consciousness taxonomy for artificial agents, which can be used as a conceptual framework for evaluating the potential level of consciousness of a given agent. Most of current implementations of artificial agents fall between levels 2 and 4 inclusive. The classification of any current implementation as fully belonging to level 5 could be thoughtfully discussed elsewhere; nonetheless, we think these kinds of agents are within current technology possibilities.

Identifying consciousness by means of interpreting behavior remains an open problem that is being currently addressed primarily in mammals, cephalopods, and birds [12, 29]. However, more effort should be put in the domain of artificial agents.

7. ACKNOWLEDGMENTS

This research has been supported by the Spanish Ministry of Education and Science under project TRA2007-67374-C02-02.

8. REFERENCES

- [1] Arp R. The Environments of Our Hominin Ancestors, Tool-usage, and Scenario Visualization. *Biology and Philosophy*, 21, 1 (2006), 95-117.
- [2] Arrabales R., Ledezma A. and Sanchis A. Modeling Consciousness for Autonomous Robot Exploration. In *IWINAC 2007*. 2007.
- [3] Baars B. J. The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Science*, 6, (2002), 47-52.
- [4] Baars B. J. *A Cognitive Theory of Consciousness*. Cambridge University Press, New York, 1993.
- [5] Bauer R. In Search of a Neuronal Signature of Consciousness – Facts, Hypotheses and Proposals. *Synthese*, 141, 2 (2004), 233-245.
- [6] Block N. On a Confusion about a Function of Consciousness. *Behav. Brain Sci.*, 18, (1995), 227-287.
- [7] Chalmers D. Consciousness and its place in nature. In Chalmers D. ed. *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press, New York, 2002.
- [8] Chalmers D. Moving forward on the problem of consciousness. *Journal of Consciousness Studies*, 4, 1 (1997), 3-46.
- [9] Ciompi L. Reflections on the role of emotions in consciousness and subjectivity, from the perspective of affect-logic. *Consciousness & Emotion*, 4, 2 (2003), 181-196.
- [10] Dennett D. C. *Consciousness Explained*. Little, Brown and Co, Boston, 1991.
- [11] Dobbyn C. and Stuart S. The Self as an Embedded Agent. *Minds and Machines*, 13, 2 (2003), 187-201.
- [12] Edelman D. B., Baars B. J. and Seth A. K. Identifying hallmarks of consciousness in non-mammalian species. *Consciousness and Cognition*, 14, 1 (3 2005), 169-187.
- [13] Gallup G. G. Self-recognition in primates: A comparative approach to the bidirectional properties of consciousness. *American Psychologist*, 32 (1977), 329-337.
- [14] Gavrillets S. and Vose A. The dynamics of Machiavellian intelligence. *PNAS*, 103, 45 (2006), 16823-16828.
- [15] Jack A. and Roepstorff A. Why Trust the Subject? *Journal of Consciousness Studies*, 10, 9-10 (2003).
- [16] Kitamura T., Otsuka Y. and Nakao T. Imitation of Animal Behavior with Use of a Model of Consciousness-Behavior Relation for a Small Robot. In *4th IEEE International Workshop on Robot and Human Communication*. Tokyo, 1995, 313-316.
- [17] Levine J. *Materialism and Qualia: The Explanatory Gap*. *Pacific Philosophical Quarterly*, 64 (1983).
- [18] Lewis M. The Emergence of Consciousness and Its Role in Human Development. *Ann NY Acad Sci*, 1001, 1 (2003), 104-133.
- [19] Manson N. State consciousness and creature consciousness: a real distinction. *Philosophical Psychology*, 13 (2000), 405-410.
- [20] Nagel T. What Is It Like To Be a Bat? *The Philosophical Review*, 83, 4 (1974), 435-450.
- [21] Nichols S. and Grantham T. Adaptive Complexity and Phenomenal Consciousness. *Philosophy of Science*, 67, 4 (2000), 648-670.
- [22] Perner J. and Lang B. Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3, 9 (1999), 337-344.
- [23] Rao A. S. and Georgeff M. P. Modeling Rational Agents within a BDI Architecture. In James A., Fikes R. and Sandewall E. eds. *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1991, 473-484.
- [24] Rao A. S., Georgeff M. P. and Sonenberg E. A. Social Plans: A Preliminary Report. In *Proceedings of the Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Elsevier Science B.V., 1992, 57-76.
- [25] Revonsuo A. and Newman J. Binding and Consciousness. *Consciousness and Cognition*, 8, 2 (1999), 123-127.
- [26] Rosenthal D. M. Metacognition and Higher-Order Thoughts. *Consciousness and Cognition*, 9, 2 (2000), 231-242.
- [27] Schacter D. L. On the relation between memory and consciousness. In *Anonymous Varieties of memory and consciousness: Essays in honor of Endel Tulving*. Erlbaum Associates, Hillsdale, NJ, 1989, 355-389.
- [28] Sergent J. and Signoret J. Implicit Access to Knowledge Derived from Unrecognized Faces in Prosopagnosia. *Cereb. Cortex*, 2, 5 (1992), 389-400.
- [29] Seth A., Baars B. and Edelman D. Criteria for consciousness in humans and other mammals. *Consciousness and Cognition*, 14, 1 (2005), 119-139.
- [30] Takeno J., Inaba K. and Suzuki T. Experiments and examination of mirror image cognition using a small robot. *CIRA 2005*, (2005), 493-498.
- [31] Turing A. *Computing Machinery and Intelligence*. *Mind*, (1950).
- [32] Wegner D. M. and Wheatley T. Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54, 7 (1999), 480-492.
- [33] Wooldridge M. *Intelligent Agents*. In Weiss G. ed. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. The MIT Press, 1999, 27-78.